**Clustering Power Consumption**

Table of Contents

[2 Introduction 3](#_Toc94350399)

[2.1 Field of Study - Objective of the study 3](#_Toc94350400)

[2.2 Problem statement 4](#_Toc94350401)

[3 Literature Review 4](#_Toc94350402)

[4 Methodology 5](#_Toc94350403)

[4.1 Data Sources 5](#_Toc94350404)

[4.2 Description of Datasets 5](#_Toc94350405)

[4.3 Techniques of Data Analysis 7](#_Toc94350406)

[4.4 Data Limitations 7](#_Toc94350407)

[4.5 Tools and Technology 7](#_Toc94350408)

[5 Results 8](#_Toc94350409)

[5.1 Visualization of Power Generation and Generation Trend 8](#_Toc94350410)

[5.2 EDA - Power Consumption 9](#_Toc94350411)

[5.3 EDA Census 10](#_Toc94350412)

[5.4 Correlation Matrix 11](#_Toc94350413)

[5.5 Clustering 12](#_Toc94350414)

[5.5.1 Distance Matrix 12](#_Toc94350415)

[5.5.2 Visualization of K-means clustering 12](#_Toc94350416)

[6 Discussion 14](#_Toc94350417)

[6.1.1 Power Generation 14](#_Toc94350418)

[6.1.2 Power Consumption 14](#_Toc94350419)

[6.1.3 Clustering 15](#_Toc94350420)

[7 Conclusion 15](#_Toc94350421)

[8 Limitation 16](#_Toc94350422)

[9 Reference/ Acknowledgement 16](#_Toc94350423)

[10 Appendix 16](#_Toc94350424)

# Introduction

On the individual country basis, India is one of the world largest electricity generators. It is one of the world’s top 5 electricity producers. Much of the India’s reliance in in terms of power generation comes from the coal power. India has installed capacity of 393 GW as of March, 2021. India has generated the total of 1383 TWh.

On one hand India is counted as one of the top producers of the electricity in the world, whereas on the other hand it has very low per capita consumption. India also has one the highest percentage of Transmission and Distribution losses.

India has a good mix in terms mode of power generation. The mode of energy generation in India is as below.

|  |  |  |
| --- | --- | --- |
| Thermal | Nuclear | Renewable |
| * Coal * Gas * Diesel | * Uranium | * Hydro * Solar |

India’s power generation is dominated by the fossil fuel which constitutes the major part of thermal energy. As per CEA by the year 2020, power is consumed mainly in five major sectors – residential, industrial, agricultural, traction. Residential, industrial, and agricultural form the major cluster in which the power is consumed. [2]

As the installed capacity and power generation has steadily increased, the industrialization has and growth of literacy and urbanization has also increased the consumption of electricity. India’s per capita consumption hovers around 1200 kWh per capita. [2] [3]

## Field of Study - Objective of the study

The objective of this project is to apply the newer tools in the field of data and business analytics to power generation and consumption and to understand any pattern.

The main idea behind this project is to understand how to apply K-means clustering techniques to power generation and power consumption by each state of India, to apply the data visualization techniques to power generation data available and provided by ministry of power and national power portal. The questions that team is to trying to explores is which are some of the census parameters which contributes to the consumption of the data. Team wanted to explore the clustering of states and union territory in India based on the power and other census attributes.

The project study is limited to 3-4 major census parameters like urbanization, literacy, and population.

## Problem statement

1. Analyze the data for electrical power generation, power consumption by states, and census.
2. Apply analytical techniques studied during the course using R
3. Perform exploratory data analytics (EDA) techniques on data
4. Attempt to run the unsupervised machine learning algorithm to cluster the states based on power consumption.
5. Infer the conclusions.

Purpose of the study is to be able to explore which census parameters have major role in electricity consumption. It is also explored using EDA, what is the energy generation mix and what are the pattern for each mode of power generation.

# Literature Review

In the past decade much of the study in the field of power sector has been devoted to plethora of subject around renewable power generation, power transmission, role of technology and internet of things using smart metering and smart electricity grids.

Some of the existing studies and application of statistical techniques using R are existing based on consumption of electricity in clusters where SC and ST population has a fair share. This project has taken an inspiration from the same but with a different perspective around literacy and urbanization. []

# Methodology

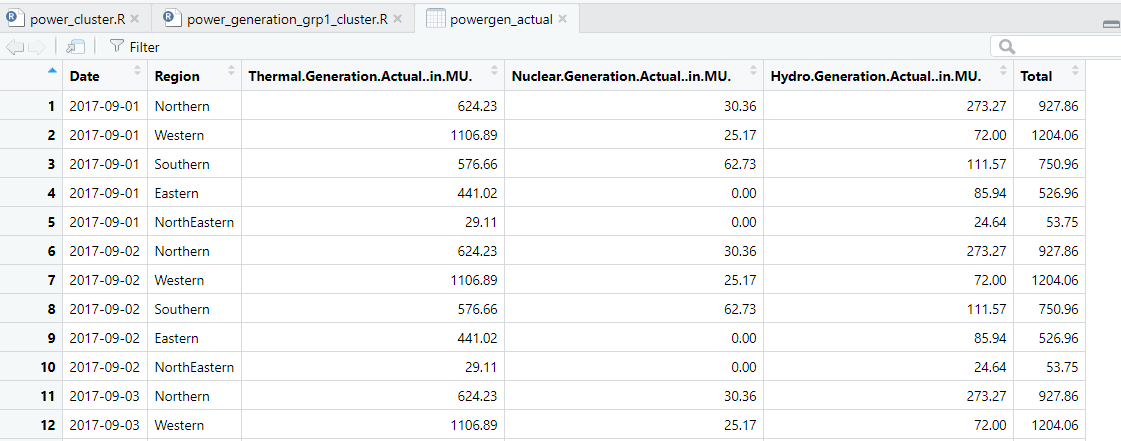
## Data Sources

The data used in the project were collected from the multiple places and primarily from the census information. The data around power generation and consumption was made available using an open-source data by national power portal <https://npp.gov.in/>). The curated data set around generation and consumption was extracted from the csv files made available by users on the Kaggle community. The data for report writing takes references from multiple datasets and report generated using information available in the report published by Central Electricity Authority in the month of October, 2020.

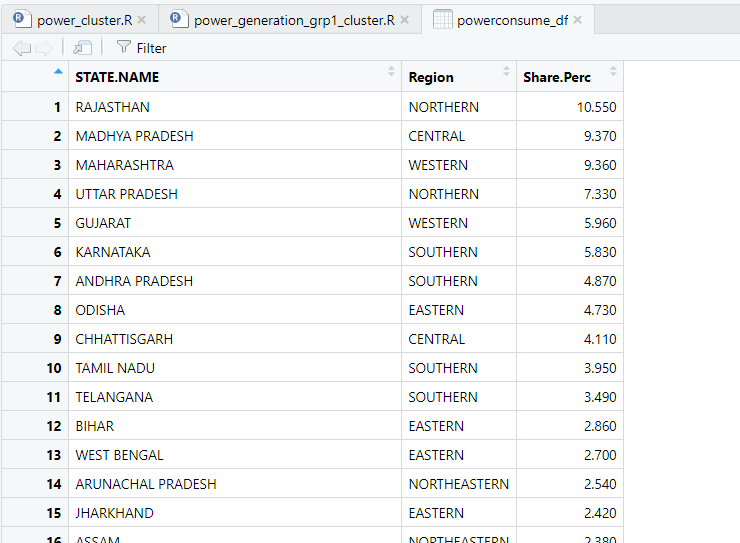
* Public information available @ National Power Portal at - <https://npp.gov.in/>
* Census Information
* <https://data.gov.in/catalog/district-wise-gdp-and-growth-rate-current-price2004-05>
* <https://data.gov.in/catalog/district-wise-gdp-and-growth-rate-constant-price1999-2000>
* <http://udise.in/drc.htm>
* Curated Region wise daily power generation – 2017-2020 and curated Elementary census data - [Exploratory Data Analysis with k-means Clustering | Kaggle](https://www.kaggle.com/aishwaryasharma1992/exploratory-data-analysis-with-k-means-clustering/data)
* india-districts-census-2011.csv - Population enumeration data with expanded columns.
* indiacensushousing-hlpca-full.csv - Housing statistics for total (rural + urban) population by district.

## Description of Datasets

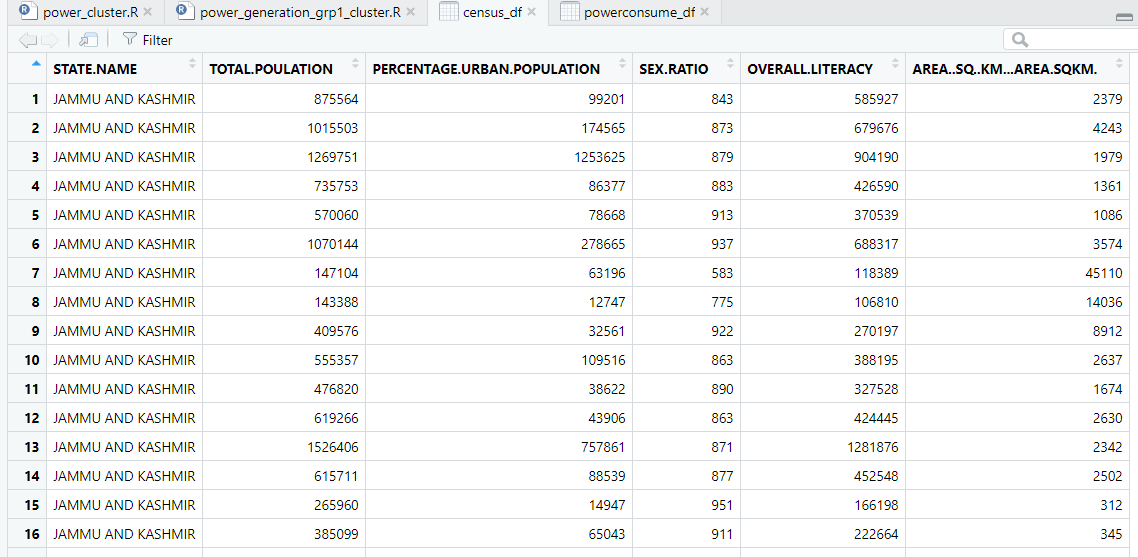
* daily\_power\_generation – Region wise and mode wise generation of power



* State\_Region\_corrected – Percentage share of electricity



* elementary\_2015\_16 – State wise and district wise parameters for population.



## Techniques of Data Analysis

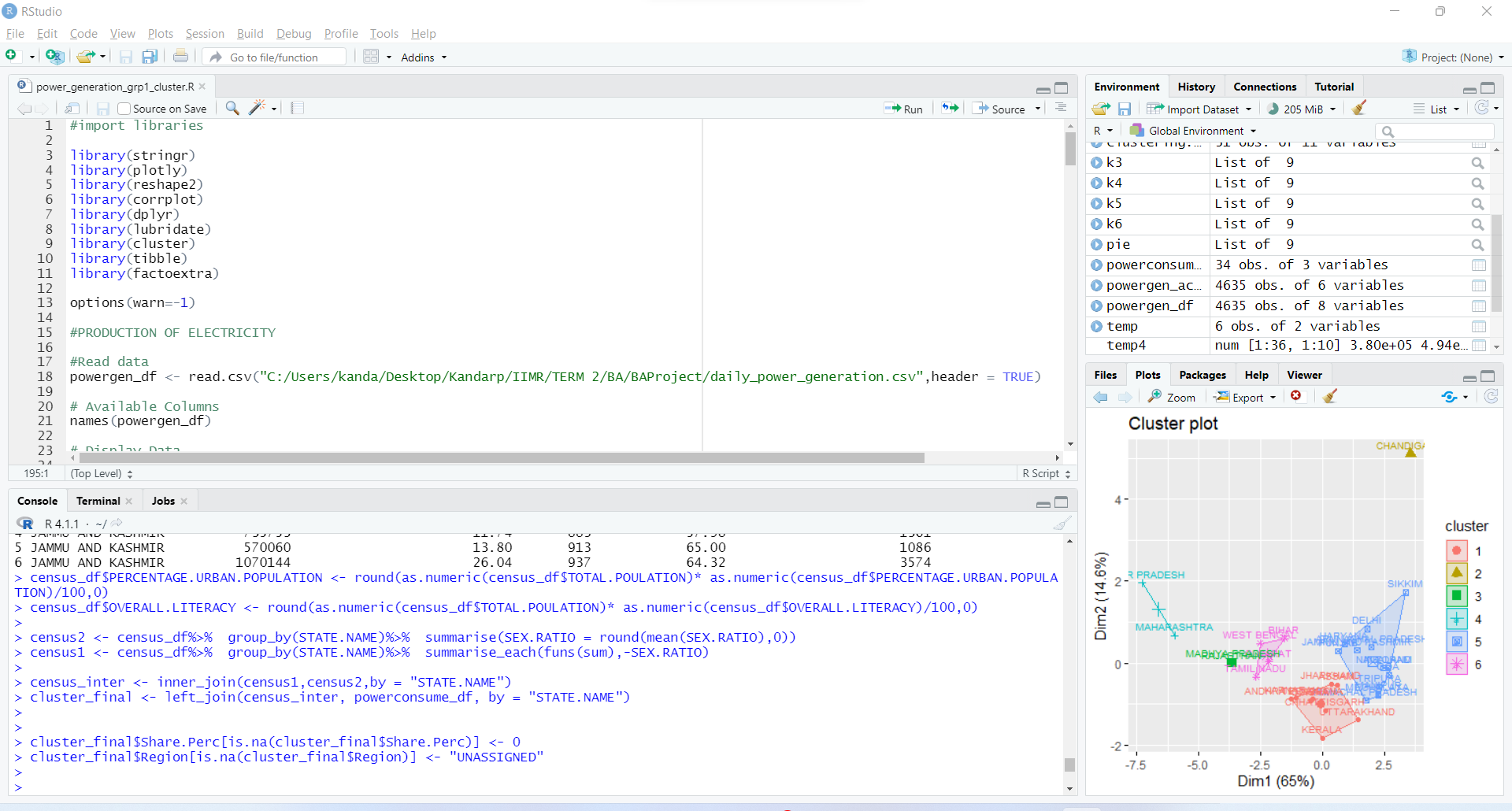
The curated data around power generation, state and region wise percentage share in the consumption, and census was imported to R notebooks and EDA applied. After the initial application of EDA, data cleaning techniques in the R were applied.

## Data Limitations

* Data used in the project was secondary data. No primary data was collected for the project. Project team did not have enough time to collect the primary data.
* The power generation dataset is curated and has no information on generation except Hydro Power. The project team could not take into consideration the same for the project.
* Power generation, power consumption data set had multiple issues of missing data values. Team has applied data filling techniques by replacing the missing values with zeroes (0) instead of mean values. It can have a bearing on the clustering of states

## Tools and Technology

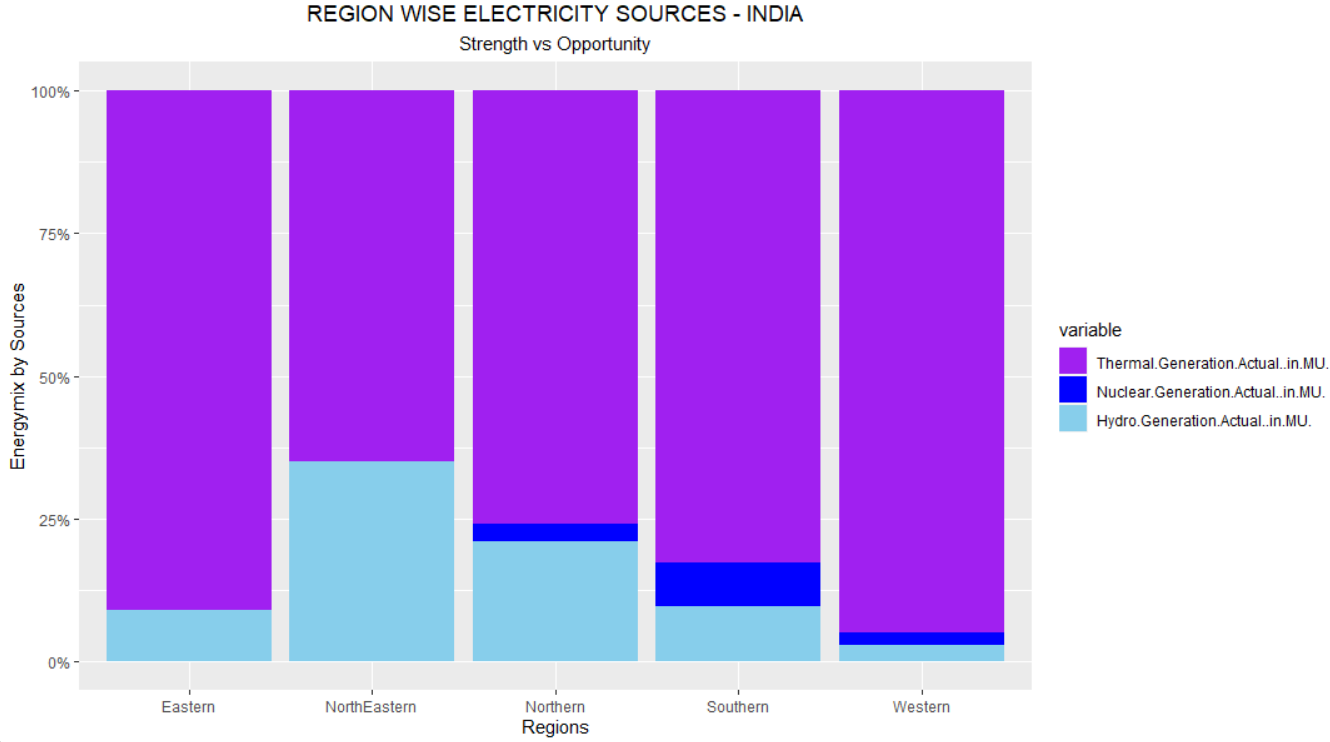
Statistical software R, and Integrated Development Environment (IDE) R-Studio has been used to perform the analytics on the data. Below is the snapshot of the same for illustration purpose.

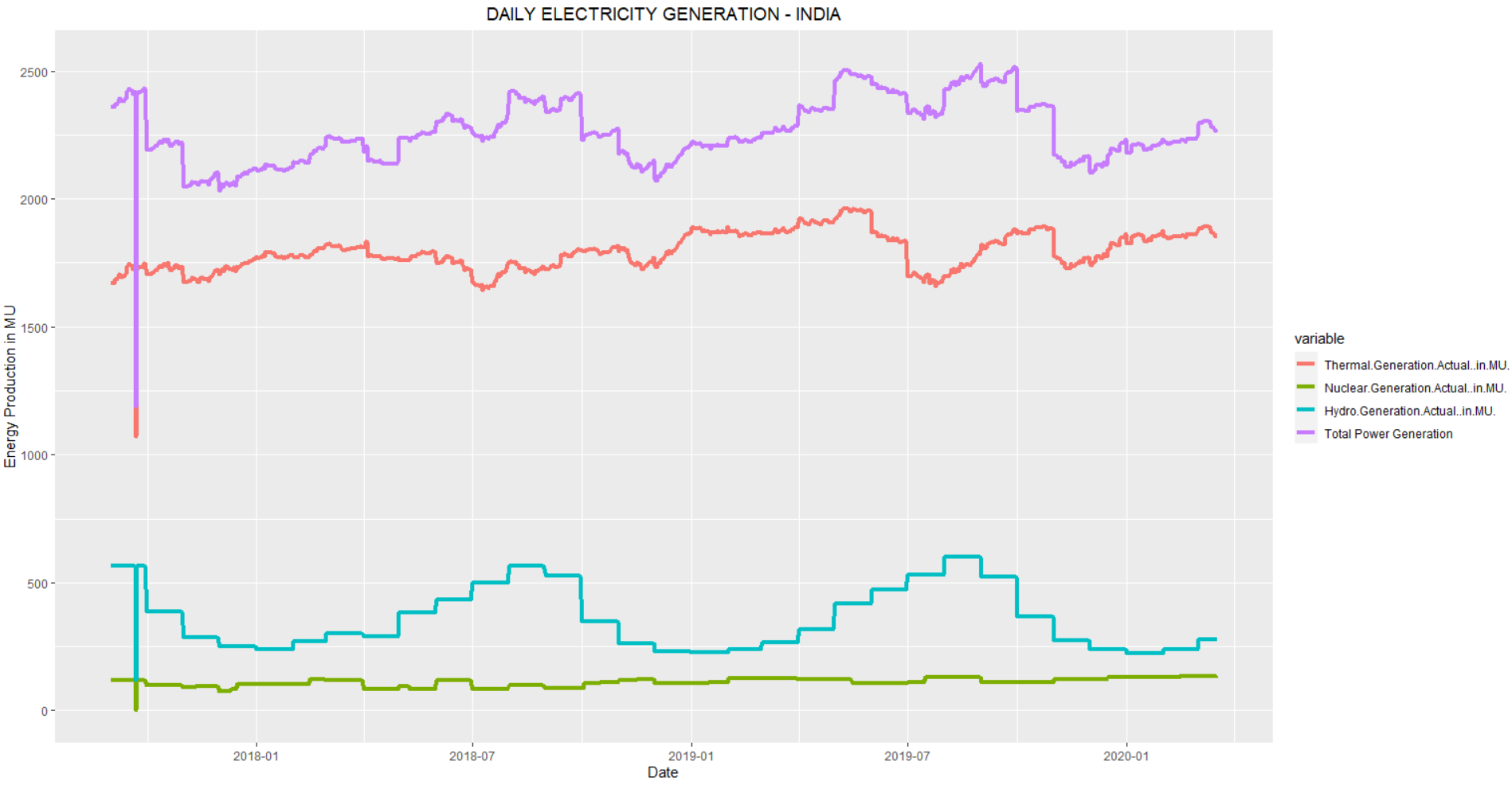


# Results

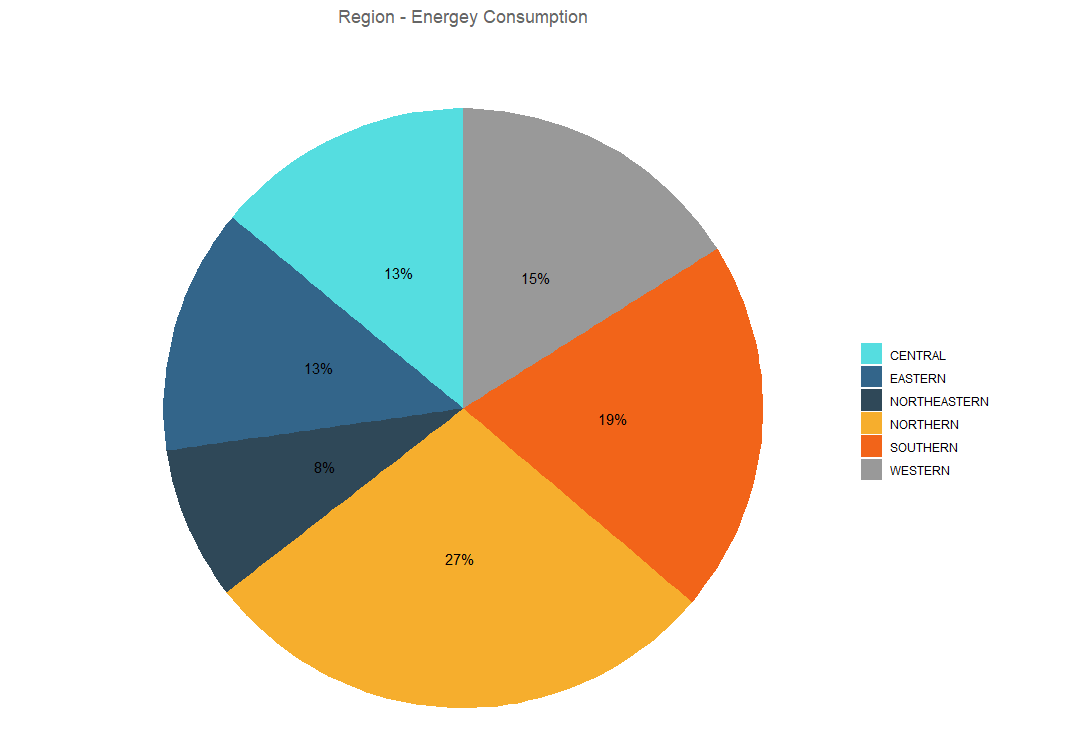
## Visualization of Power Generation and Generation Trend

Below two visualizations performed on the power generation throws some light what is the mix of mode of power generation.





## EDA - Power Consumption



The pie chart above shows that the Northern and Southern region of the country has approximately 46% share in the power consumption.

* What are the contributing factors of power consumption compared to other regions?
* What are the census parameters that contributes to power consumption?

In the sections below of this project, team will try to attempt to answer these questions.

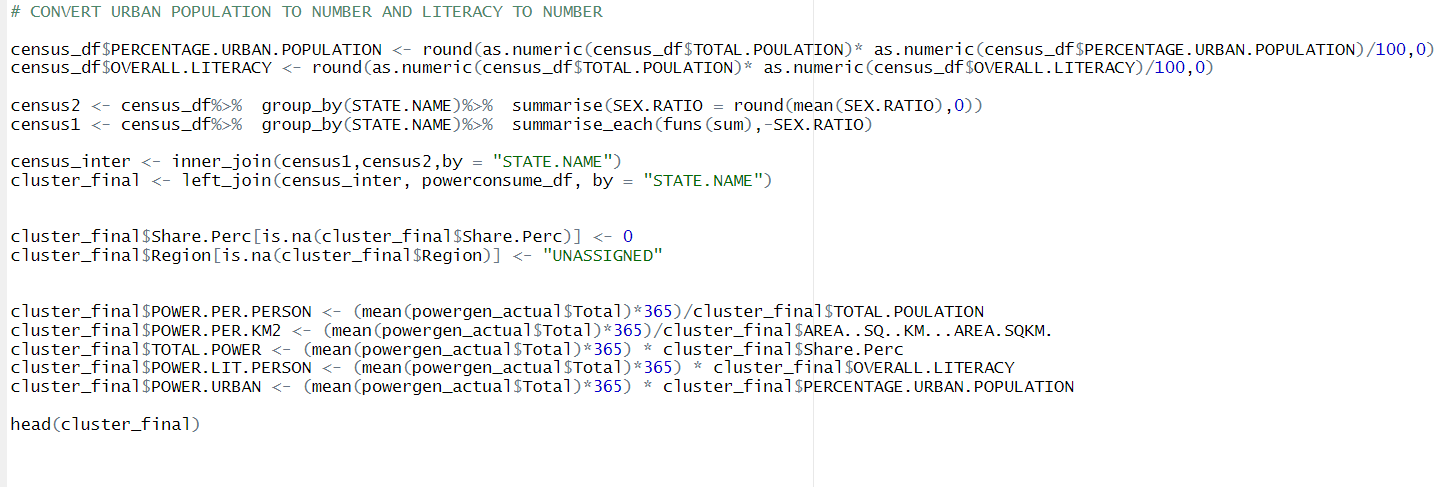
## EDA Census

Based on the EDA for the curated census data few variables are primarily used to cluster the total consumption of power. Out of these 5 variables Power consumption per person,

* Total power
* Power per square kilometers
* Power consumption per person
* Power per literate person
* Power per urban population

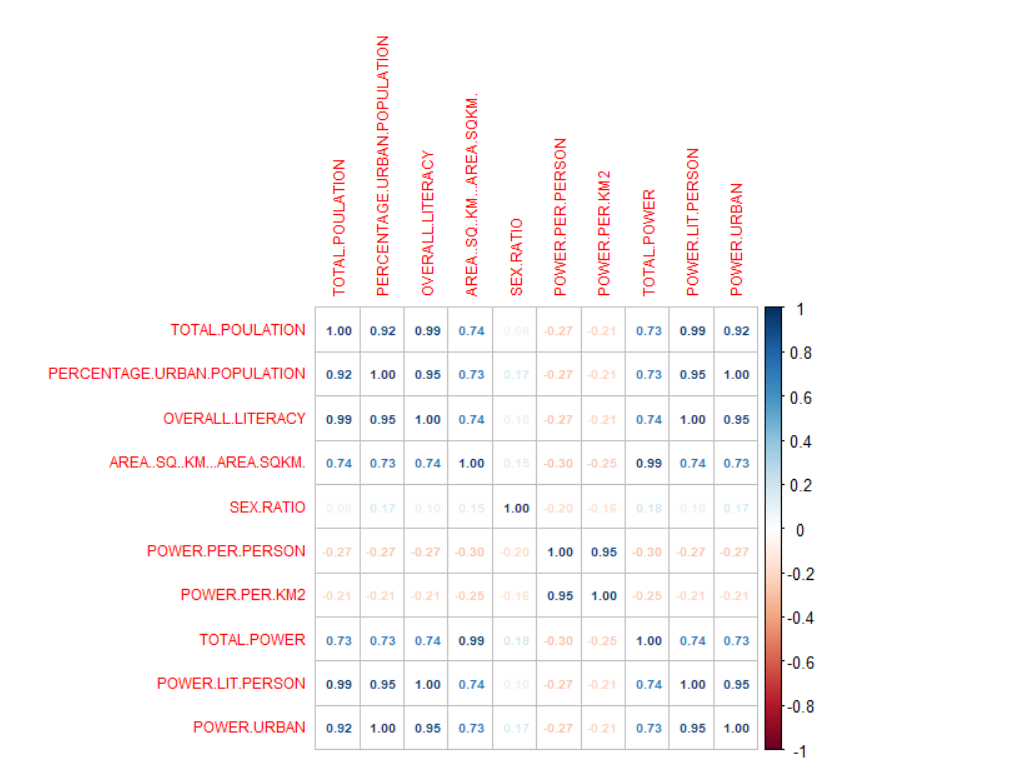
Industry uses the PCA (Principal Component Analysis) or Dimensionality reduction to finalize the important variables from amongst the hundreds of variables. Based on the eigen vector and eigen value analysis to reduce the number of variables.

Group did not have much time to employ this technique. So, team selected the variables which based on common knowledge.



## Correlation Matrix

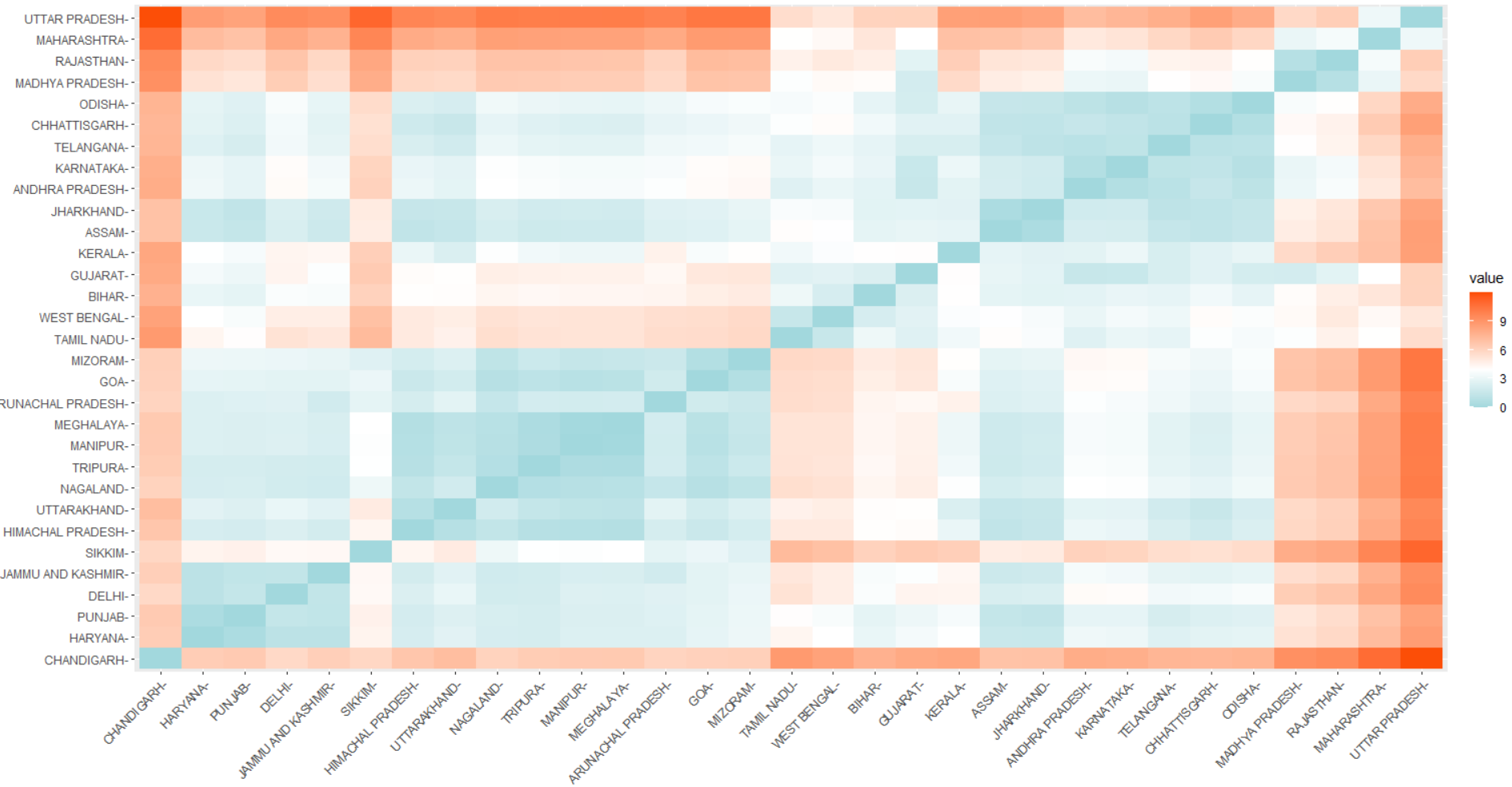
It is visible from the correlation matrix that percentage of the urban population, literacy and total population are highly correlated.



## Clustering

### Distance Matrix

Applying the distance matrix on the final dataset for clustering.

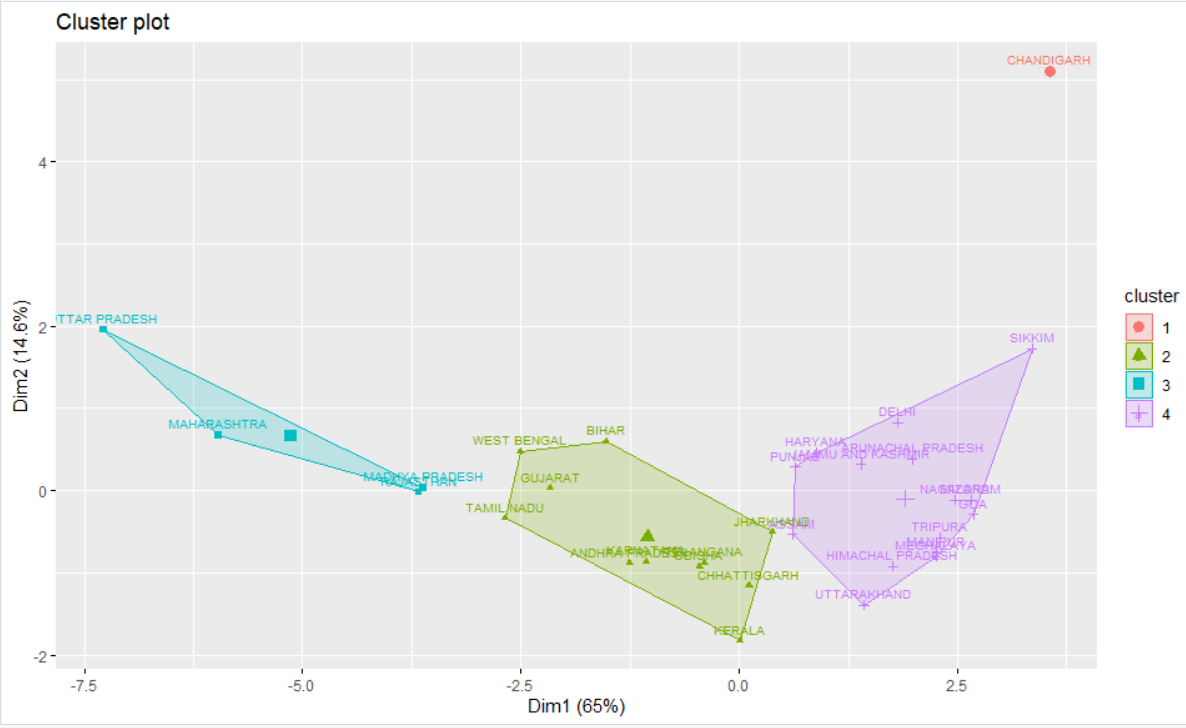


### Visualization of K-means clustering

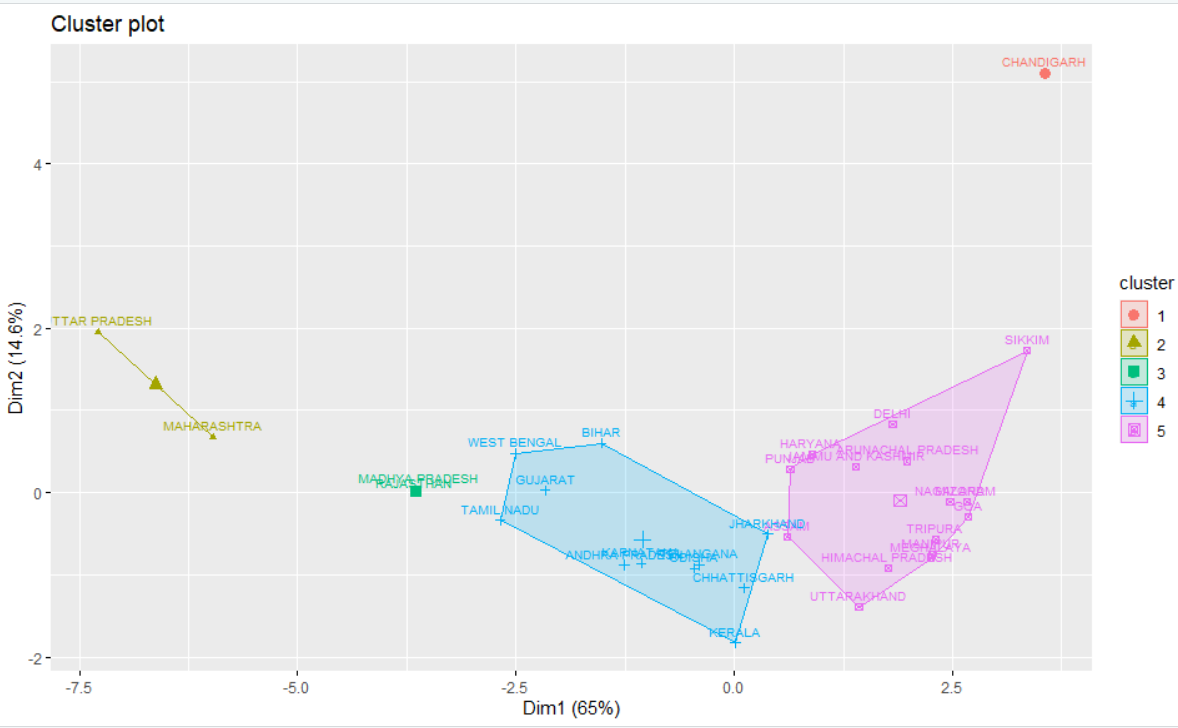
Cluster plot where K = 3



Cluster plot where k=4



Cluster plot where k=5



# Discussion

### Power Generation

* Thermal power still constitutes the majority in power generation. Government is systematically working towards promoting alternate source of energy.
* North East and Norhtern regions have good presence of Hydro power. Eastern region through its terrain and northern region through its multiple dams present on rivers contributes to majority of hydro power.
* Power generation trends conveys that total power generation trend matches the thermal power generation trend due to thermal being the major source of energy. It has more weightage.
* Nuclear power generation has not increased over the period of 3 years of 2017-2020.
* Hydro power follows the pattern of monsoon – that is why the graph is cyclic.

### Power Consumption

* Northern and Southern region contributes majorly to energy consumption

### Clustering

* When K = 3, the cluster includes outliers like Chandigarh. It doesn’t represent the correct picture.
* K=4, most of the states are grouped based on common characters and Chandigarh is clearly visible as outlier.
* K=5, states are again fragmented but not much difference in the selected variable like population and literacy between groups.
* K = 4 clusters the states with most similarity and is apt for this analysis.
* Total Population, Urban Population and Literacy are the major factors contributing the consumption power.

# Conclusion

* Literacy and urbanization impact the consumption of power. Having a cluster of Maharashtra, Uttar Pradesh in the same cluster confirm this to an extent.
* Maharashtra is in a different cluster than both Gujarat and Tamil. Gujarat and Tamil Nadu are in the same cluster. Maharashtra, Gujarat, and Tamil Nadu are highly industrialized. But since we have no data available for the industry. Clustering does not throw any light on this aspect.
* Thermal power (biofuel) is still the major factor in power generation, in the future the energy from the cleaner source of the energy can be promoted and looked at.
* Hydro power follows the weather pattern. It is cyclic and follows pattern of river water inflow. If data for solar power is available it can there can be a comparative study for the same.
* Nuclear power generation is still miniscule compare to other mode of power generation and has not grown in the period of three years under study. i.e. 2017-2020.

# Limitation

* The power sector has very large and any one study or project can not throw light on multiple aspects.
* There can be some missing factors while doing clustering in this project. E.g. losses in Transmission and distribution can affect the consumption of power or how much demand was not satisfied during peak season.
* Other studies which are more research oriented can throw more light on the other aspects of power sector in India.

# Reference/ Acknowledgement

[1] <https://cea.nic.in/wp-content/uploads/installed/2021/12/installed_capacity.pdf>

[2] <https://cea.nic.in/wp-content/uploads/pdm/2020/12/growth_2020.pdf>

[3] Statistical Review of world Energy – 2021 - <http://www.indiaenvironmentportal.org.in/files/file/bp%20statistical%20review%20of%20world%20energy%202021.pdf>

[4] Curated Region wise daily power generation – 2017-2020 and curated Elementary census data - [Exploratory Data Analysis with k-means Clustering | Kaggle](https://www.kaggle.com/aishwaryasharma1992/exploratory-data-analysis-with-k-means-clustering/data)

# Appendix

* Data and R file



* Presentation

